Batch jobs on ClueDØ and CAB

Roger Moore

Michigan State University

What we will Cover

- When to use batch jobs
- Whether to run on CAB or CluedØ
- Scheduling policies
  - How to maximize the chance of a job running quickly
- Writing your own Batch script
  - Getting your inputs and writing your outputs
  - Differences between CAB and CluedØ
- Using mc_runjob to for Batch jobs
- Submitting batch jobs
- Low level interface to SAM
- Will not cover dØtools

Why we Need a Batch System

- Batch systems provide easy access to all the CPUs in a cluster, harnessing them to do something useful than running screensavers
  - ClueDØ usually runs ~150 jobs 24 hours/day
- They divide an entire cluster's CPU using preset rules
  - Ensures everyone gets a fairshare of the resources
- Match best available resource for the job
  - Ensures there is sufficient memory
  - Picks best available CPU
  - Stops two processes competing for 1 CPU
- Without a Batch system chaos would reign!

When To Use Batch Jobs

- Batch jobs best for processes needing moderate amounts of CPU (>~20-30 mins)
  - Below this batch system startup overhead (~1-2 mins if CPU free) will be a large fraction of time
- For repetetive, short tasks use an interactive batch job
  - Gives you a shell on a free machine and reserves the CPU for your use
- On CAB you have to use Batch jobs
  - You can login to the nodes but only to check on your batch jobs
- On ClueDØ jobs >30 mins must run in batch
  - Please obey this, we don't want to enforce limits

CAB or CluedØ?

- CAB really intended for SAM access to data
  - High bandwidth access to disks on dØmino
  - About twice the "effective" CPU power of CluedØ
    - 160 dual 1.6GHz nodes with 1GB memory each
    - Run 320 jobs concurrently vs. CluedØ's ~150 (memory limited, not CPU)
- CluedØ has SAM access but limited bandwidth
  - Especially true in DAB!
- For large I/O jobs with SAM use CAB
  - This is why it was built!

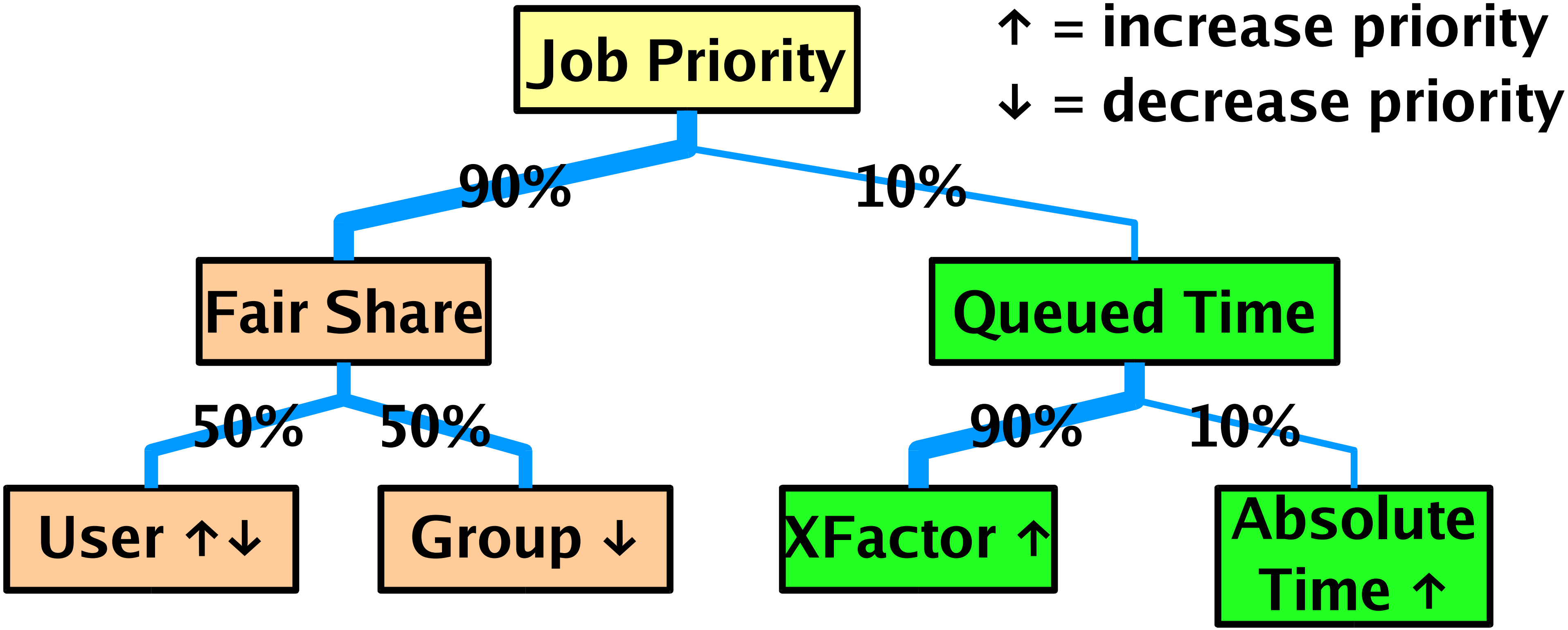
CAB or ClueDØ?

- ClueDØ best for MC or local analysis of datasets on ClueDØ disk
  - Be careful though, 50 jobs reading or writing to one disk can take out a machine!
  - Try to use disks on 1Gb connected machines for heavy I/O or use SAM
- However, ClueDØ often over subscribed and CAB half empty!
  - If so, use CAB 'medium' for MC or local analysis
  - Don't do this if CAB is full of SAM jobs though

Scheduling Policies

- CAB's very simple: first come first served
- ClueDØ's more complex!
  - Admins recently agreed to a change
  - New policy outlined here
- ClueDØ run as an institute based resource
  - Most cluster-wide resources supplied by DØ
  - Not controlled by ORB
  - No real support for physics group resources other than disk i.e. no top, NP etc. based queues
- ClueDØ system must take account of relative institute contributions
  - Encourages institutes to buy CPU when needed

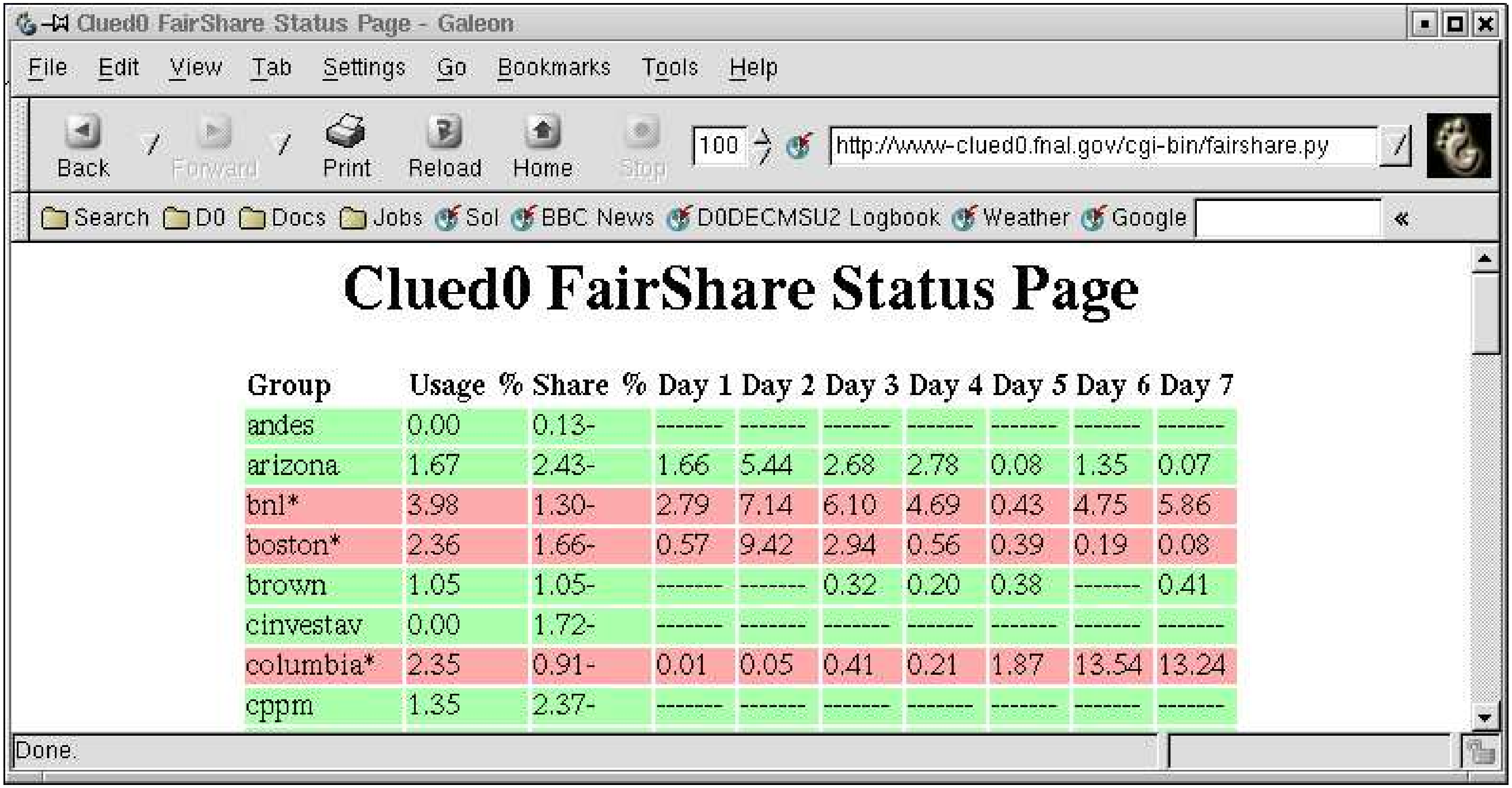
CluedØ Scheduling Policies



- *XFactor = queued time/requested CPU time*
- *Group quota = group CPU MHz/Total CPU MHz*
- *User quota = Group quota*
  - *Users below quota INCREASE job priority*

CluedØ Scheduling Policies

- *Current state of Clued0 fair share available on the web*
  - *http://www-clued0.fnal.gov/cgi-bin/fairshare.py*

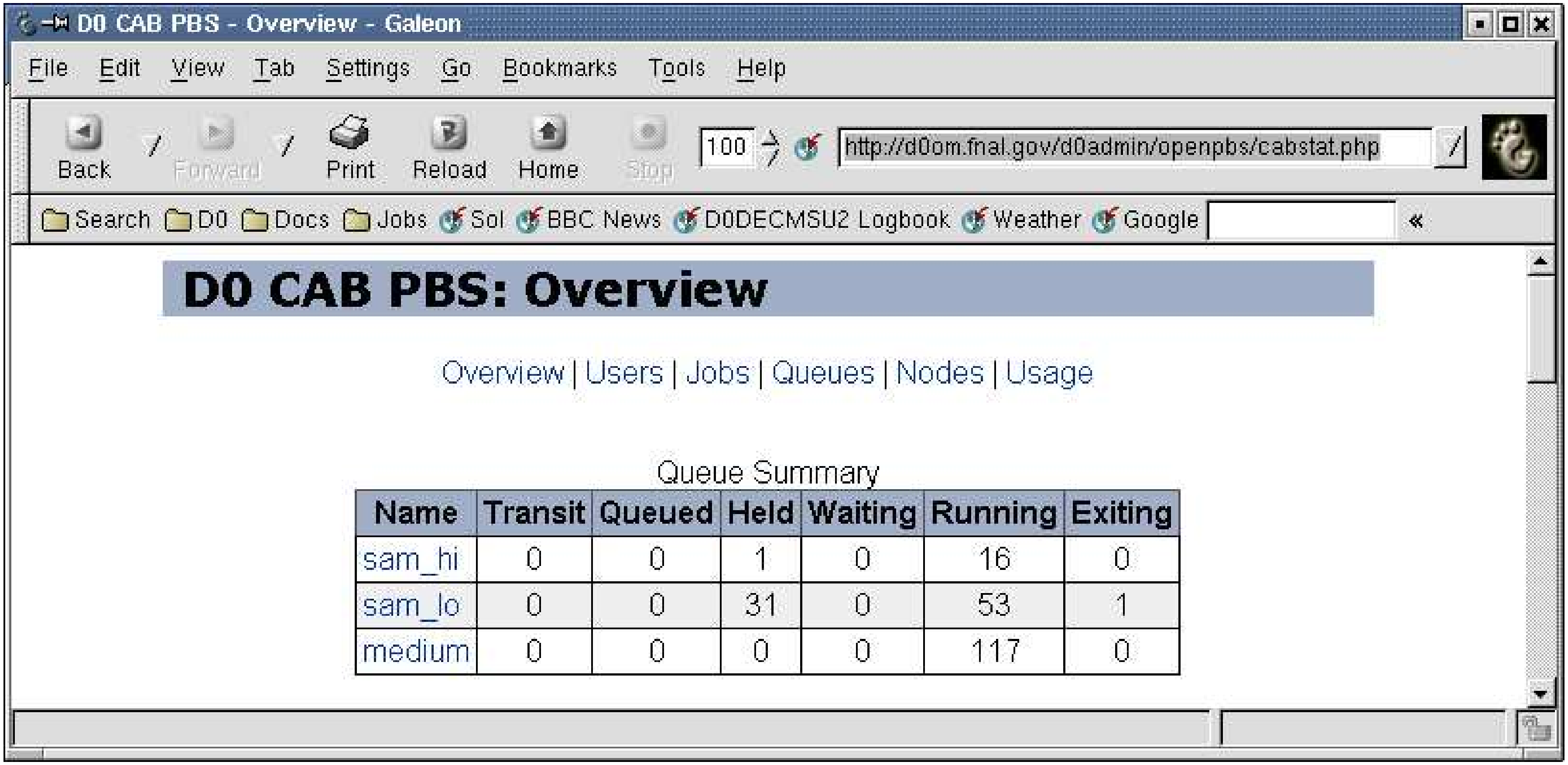

Clued0 FairShare Status Page - Galeon
File Edit View Tab Settings Go Bookmarks Tools Help
Back
Forward
Print
Reload
Home
Stop
100
http://www-clued0.fnal.gov/cgi-bin/fairshare.py
Search D0 Docs Jobs Sol BBC News D0DECMSU2 Logbook Weather Google
Done.

Clued0 FairShare Status Page

Group	Usage %	Share %	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
andes	0.00	0.13-	-------	-------	-------	-------	-------	-------	-------
arizona	1.67	2.43-	1.66	5.44	2.68	2.78	0.08	1.35	0.07
bnl*	3.98	1.30-	2.79	7.14	6.10	4.69	0.43	4.75	5.86
boston*	2.36	1.66-	0.57	9.42	2.94	0.56	0.39	0.19	0.08
brown	1.05	1.05-	-------	-------	0.32	0.20	0.38	-------	0.41
cinvestav	0.00	1.72-	-------	-------	-------	-------	-------	-------	-------
columbia*	2.35	0.91-	0.01	0.05	0.41	0.21	1.87	13.54	13.24
cppm	1.35	2.37-	-------	-------	-------	-------	-------	-------	-------

CAB Batch Status

- CAB also has a very nice webpage with their batch system status on display
  - http://d0om.fnal.gov/d0admin/openpbs/cabstat.php


D0 CAB PBS - Overview - Galeon
File Edit View Tab Settings Go Bookmarks Tools Help
Back Forward Print Reload Home Stop
100
http://d0om.fnal.gov/d0admin/openpbs/cabstat.php
Search D0 Docs Jobs Sol BBC News D0DECMSU2 Logbook Weather Google

D0 CAB PBS: Overview

Overview | Users | Jobs | Queues | Nodes | Usage

Queue Summary

Name	Transit	Queued	Held	Waiting	Running	Exiting
sam_hi	0	0	1	0	16	0
sam_lo	0	0	31	0	53	1
medium	0	0	0	0	117	0

Other Factors

- Scheduling only determines relative priority
- CluedØ not CPU limited: memory limited!
  - Always (so far) idle CPUs with little memory
- Batch system reserves 128MB for desktop and ~12 MB for kernel
  - Available memory is 140MB less than installed
- Memory quantized: threshold effects
  - Single CPU machines: 116MB, 372MB, 884MB
  - Dual CPU machines: 58MB, 186MB, 442MB
- If you can get you job below one of these thresholds you will increase the number of machines it can run on!

Writing a Batch Script

- Batch script needs to do three things
  - Copy files needed to run the job to the local scratch area
  - Run the executable(s)
  - Copy the results back to a given location
- CluedØ and CAB have different filesystems
  - CAB has access to dØmino project disks
  - CluedØ has access to /work and /rooms
- CAB has 'kbatch' script to get a Kerberos ticket for use with batch scripts
  - Stores a kerberos password in a file!
  - Asked not to implement this for CluedØ at moment

Making 'kbatch' Work

- 'kbatch' needs one-time setup
- Login to dØmino and type the following:

```
> setup kcroninit
> kcroninit
```

- This creates a new Kerberos principle
- Change your dØmino and CluedØ ~/.k5login files to read

```
<username>@FNAL.GOV
<username>/cron/d0mino.fnal.gov@FNAL.GOV
```

- IMPORTANT: There must be no spaces at the end of lines in the .k5login!

Detecting where you run

- *Need to customize depending on machine*

```
echo $HOST | grep clued0
if [ $? ]; then # NOT a clued0 machine
    ...
else            # Clued0 machine
    ...
fi
```

- *Could use this to set custom variables:*

```
CLUEDOHOME=$HOME
DOMINOHOME=/d0mino/$USER
```

  - *..or in CAB's case:*

```
CLUEDOHOME=$HOME/desktop
DOMINOHOME=$HOME
```

- *and also useful for changing how you get inputs or write outputs*

Using Local Scratch Area

- Both CAB and Clued0 provide a private scratch directory for the job
  - Destroyed when job completes so be careful to copy out everything you need
  - Improves job speed (and reliability) considerably if you run from local directory
- Path differs for two systems
  - CAB: /scratch/$PBS_JOBID
  - CluedØ: /batch/$PBS_JOBID
- Copy files to this directory, run the executables and then copy back

Example: mc_runjob & CAB

- Script used to run mc_runjob jobs on CAB
  - Very similar to one generated by mc_runjob
  - Useful for customizing to run your own jobs
    - e.g. ROOT based analyses
- First get initial environment and chose a destination directory

```
#!/usr/bin/env bash
# This gets the environment
. /etc/bashrc
. /usr/products/etc/setups.sh

DESTDIR=.../mc/bbbar-incl/10GeV/$PBS_JOBID
export DESTDIR
```

Example: mc_runjob & CAB

- *Change to the scratch directory, get a kerberos ticket and copy over the initial files*

```
cd /scratch/$PBS_JOBID
kbatch

setup D0RunII p13.08.00
setup -t mc_runjob
rcp thwaite:/.../bbbar-incl-10GeV.macro .
mc_runjob -macro=bbbar-incl-10GeV.macro
```

- *Now actually execute the job*

```
mc_runjob -macro=bbbar-incl-10GeV.macro
```

Example: mc_runjob & CAB

- When processing complete get a new ticket and copy the output back

```
kbatch
rsh ripon-clued0 mkdir -p $DESTDIR
tar zcf - * | rsh -X ripon-clued0 \
    "cd $DESTDIR;tar zxf -"
# Job complete!
```

- Complex copy back command needed
  - Avoids converting symbolic links into real files
  - Compresses output to take less bandwidth
    - CPU not a problem on CAB
  - Turns off kerberos encryption
- Use "mc_jobscript" command to auto-generate your own

Submitting a Job

- So now you have written your job you need to submit it
- Clued0 supplies two different commands
  - cluesow: sends your job to the Clued0 queues
  - cabsow: sends you job to the CAB batch queues
- Syntax same for both:

```
    > ...sow -q <queue> <script>
```

- For CAB no need to specify memory or CPU time: no limits on either!
- For Clued0 need to add:

```
    -l cput=<h>:<m>:<s> -l mem=<X>mb
```

Submitting a Job

- Clued0 has "wall time" limits too
  - Default is 3 x default CPU
  - Removes hung jobs, special reservations
- Clued0 queues:
  - SHORT: cpu<3 hours, wall<6 hours, 2GB [5 slt/day]
  - MEDIUM: cpu<12 hrs, wall<36 hrs, 1GB
  - LONG: cpu<72 hrs, wall<144 hrs, 1GB
  - FAST: cpu<12 hrs, wall<12 hrs, 1GB [5 slots/night]
- CAB queues:
  - cabmed: All non-SAM jobs
  - cab: All SAM related jobs

Low Level SAM Interface

- Use SAM interface built into executable
  - More flexibility for adding "make up" jobs
  - Gives rough handle on CPU time/job
    - Essential for Clued0 with batch CPU time limits
- Start a SAM project manually
  - sam start project.....
- In the batch script which runs a d0 executable add the command line options
  - d0exe --project=... --num_files=<max>
- Now just submit your script to the batch system enough times to process all the files

Low Level SAM Interface

- If any of your jobs crash just submit each ones to make up the slack
  - Apparently possible to recover non-processed files but not easy
- Once all the jobs are complete run:
  - sam stop project --name=<name>
- ...and that's it!

(Ab)Using the Clued0 System

- Rule of thumb for Clued0 system:
  - Unless you deliberately try, almost nothing you submit will abuse the batch system
  - This is NOT a challenge! Just a statement not to be too shy...please check if you think trouble likely!
  - Two exceptions: be careful of I/O to single disk, don't write 1+GB to stdout or stderr!
- The scheduler will stop you hogging all the cluster resources
  - Your institutional colleagues might not thank you for using their group quota to run 500 jobs of SETI@home but that's not a Clued0 problem (as long as you run them in the batch system!)

(Ab)Using the Clued0 System

- If you are going to submit > 500 jobs at once
  - So far this is not a common occurance so please notify clued0-admin so we know something hasn't gone nuts!
  - Use the 'sleep' command to put ~10s between submits
- One user successfully submitted and ran a 2000 job project in a day!
- Don't worry about when you submit jobs, Clued0 will queue them until it can run them
  - Higher priority the longer a job waits
  - No limit to the number of jobs queued (unlike CAB)

Conclusions

- You should now be able to write and submit your own batch scripts
- Generally a good idea to run small test jobs with new scripts before submitting lots
- If you have problems contacting Clued0 batch system then wait ~5 minutes
  - Script automatically restarts it
  - If all else fails contact clued0-admin@fnal.gov
  - No 24 hour support (but admins in Europe!)
- CAB also has auto-restarter
  - Contact address is d0cab-users@fnal.gov
  - 24 hour support through helpdesk